



**NETWORK INTRUSION DETECTION**

**WITH**

**NAÏVE BAYES CLASSIFICATION AND SELF**

**ORGANIZING MAPS**

Master's Student: Mubeen Iqbal

Supervised by: A/Prof. Quang Ha

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY (FEIT)

UNIVERSITY OF TECHNOLOGY SYDNEY (UTS)

August 2014

## **CERTIFICATE OF ORIGINAL AUTHORSHIP**

*I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.*

*I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.*

*Mubeen Iqbal*

## **Acknowledgment**

First and foremost, I offer my sincerest gratitude to my advisor, **A/Prof. Quang Ha**, who has backed me, all around my thesis with his perseverance and knowledge. This research would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

Secondly, I would like to thank my research colleagues for their encouragement throughout my research work.

Finally, I would like to thank my parents for their unending love, support and understanding during my research work.

# Table of Contents

**Acknowledgments**

**Certificate of Original Authorship**

**Table of Contents**

**List of Tables**

**List of Figures**

**List of Abbreviations**

**Abstract**

|  |    |
|--|----|
| CHAPTER 1 .....  | 12 |
| INTRODUCTION .....   | 12 |
| 1.1. Overview .....  | 12 |
| 1.2. Network Intrusion .....   | 12 |
| 1.3. Network Intrusion Detection.....  | 13 |
| 1.4. Problem Statement .....   | 14 |
| 1.5. Objectives .....  | 15 |
| 1.6. Scope of Thesis .....   | 15 |
| 1.7. Organization of the Thesis .....  | 16 |
| CHAPTER 2 .....  | 17 |
| INTRUSION DETECTION SYSTEMS.....   | 17 |
| 2.1. Introduction.....   | 17 |
| 2.2. Computer Security and its Role .....                                    | 17 |
| 2.2.1. Threats to Security .....   | 18 |
| 2.2.2. Identifying Threats.....  | 19 |
| 2.3 Diagrammatic Representation of PC Frameworks with and without IDS: ..... | 21 |
| 2.3.1 Scenario I .....   | 21 |
| 2.3.2 Scenario II.....   | 21 |
| 2.4 What is Intrusion Detection?.....  | 22 |
| 2.4.1 Introduction.....  | 22 |
| 2.4.2. Network Intrusion Detection Systems .....                             | 23 |
| 2.4.3. Evolution of Intrusion Detection.....                                 | 23 |
| 2.4.3 IDS Components .....   | 30 |
| 2.4.4 IDS Methodologies .....  | 34 |
| 2.4.5 Types of IDS .....   | 35 |

|                                   |   |    |
|-----------------------------------|---|----|
| 2.4.6                             | Applications of Network Intrusion Detection Systems.....        | 36 |
| 2.5                               | Summary .....   | 40 |
| CHAPTER 3 .....                   |   | 42 |
| BACKGROUND AND RELATED WORK ..... |   | 42 |
| 3.1                               | Introduction.....   | 42 |
| 3.2                               | Anomaly Intrusion Detection.....                                | 43 |
| 3.2.1                             | Statistical Technique .....                                     | 45 |
| 3.2.2                             | Feature Selection.....  | 48 |
| 3.2.3                             | Predictive Pattern Generation .....                             | 49 |
| 3.2.4                             | Neural Networks .....   | 50 |
| 3.2.5                             | Bayesian Classification.....                                    | 52 |
| 3.2.6                             | Belief Networks .....   | 53 |
| 3.3                               | Misuse Intrusion Detection .....                                | 54 |
| 3.3.1                             | Using Conditional Probability to Predict Misuse Intrusions..... | 55 |
| 3.3.2                             | Production/Expert Systems in Intrusion Detection .....          | 56 |
| 3.3.3                             | State Transition Analysis .....                                 | 57 |
| 3.3.4                             | Keystroke Monitoring .....                                      | 57 |
| 3.3.5                             | Model-Based Intrusion Detection .....                           | 58 |
| 3.4                               | A Generic Intrusion Detection Model.....                        | 59 |
| 3.5                               | Comparison with Other Systems.....                              | 61 |
| 3.6                               | Shortcomings of Current Intrusion Detection Systems.....        | 62 |
| 3.7                               | Summary of Intrusion Detection Techniques .....                 | 63 |
| CHAPTER 4 .....                   |   | 66 |
| MACHINE LEARNING APPROACHES.....  |   | 66 |
| 4.1                               | Introduction.....   | 66 |
| 4.1.1                             | Definition .....  | 67 |
| 4.1.2                             | Formal Description and Terminology .....                        | 68 |
| 4.1.3                             | Application.....  | 69 |
| 4.1.4                             | Benefits of Machine Learning.....                               | 69 |
| 4.2                               | Supervised Versus Unsupervised Learning .....                   | 70 |
| 4.3                               | Supervised Learning .....                                       | 72 |
| 4.3.1                             | Naïve Bayes .....   | 77 |
| 4.3.2                             | Linear Discrimination Analysis (LDA) .....                      | 84 |
| 4.3.3                             | Artificial Neural Networks.....                                 | 87 |
| 4.3.4                             | Support Vector Machines (SVMs).....                             | 88 |
| 4.4                               | Unsupervised Learning .....                                     | 92 |

|   |                                    |     |
|---|------------------------------------|-----|
| 4.4.1   | Cluster Analysis.....              | 93  |
| 4.4.2   | Clustering Algorithms.....         | 97  |
| 4.5   | Hidden Markov model.....           | 102 |
| 4.6   | Self-Organizing Maps.....          | 104 |
| 4.6.1   | Learning Rule for SOM.....         | 105 |
| 4.6.2   | SOM Algorithm.....                 | 109 |
| 4.6.3   | SOM as Clump Technique.....        | 113 |
| 4.6.4   | SOM as Image Technique.....        | 117 |
| 4.6.5   | U-matrix.....                      | 118 |
| 4.6.6   | Component Plane Visualization..... | 121 |
| 4.7   | Summary.....                       | 121 |
| CHAPTER 5 .....                               |                                    | 125 |
| KDD CUP' 1999 DATASET.....                    |                                    | 125 |
| 5.1   | Introduction.....                  | 125 |
| 5.2   | KDD Cup99 Methodology.....         | 128 |
| 5.3   | NSL-KDD Data Set .....             | 128 |
| 5.4   | Attributes in KDD CUP99 .....      | 129 |
| 5.4.1   | Basic Features .....               | 130 |
| 5.4.2   | Content Features .....             | 130 |
| 5.5   | Symbolic Features.....             | 133 |
| 5.6   | Numeric features.....              | 136 |
| 5.7   | Classification of Attacks.....     | 140 |
| 5.7.1   | Denial of Service.....             | 142 |
| 5.7.2   | User to Root .....                 | 143 |
| 5.7.3   | Remote to User.....                | 143 |
| 5.7.4   | Probing.....                       | 143 |
| 5.8   | Summary .....                      | 145 |
| CHAPTER 6 .....                               |                                    | 147 |
| IMPLEMENTATION, EXPERIMENTS AND RESULTS ..... |                                    | 147 |
| 6.1   | Collection of the Dataset.....     | 148 |
| 6.2   | Data Processing.....               | 149 |
| 6.2.1   | Indicator Variables.....           | 150 |
| 6.2.2   | Conditional Probability.....       | 152 |
| 6.3   | Training and Testing Phase.....    | 154 |
| 6.3.1   | Self-Organizing Map.....           | 154 |
| 6.3.2   | Naïve Bayes .....                  | 159 |

|                 |   |     |
|-----------------|---|-----|
| 6.4             | Performance Evaluation.....                                   | 165 |
| 6.5             | Experiment.....   | 166 |
| 6.6             | Environment.....  | 167 |
| 6.7             | Results.....  | 167 |
| 6.8             | SOM Trained on Conditional Probability Processed Dataset..... | 170 |
| 6.9             | SOM Trained on Indicator Variable Processed Dataset .....     | 172 |
| 6.10            | Summary .....   | 173 |
| CHAPTER 7 ..... |   | 176 |
| CONCLUSION..... |   | 176 |
| 7.1             | Contribution of Thesis .....                                  | 177 |
| 7.2             | Future Work.....  | 178 |
| References..... |   | 179 |

## List of Figures

|   |     |
|---|-----|
| <i>Figure 2.1 - Unsecured PC Framework representation</i> .....   | 21  |
| <i>Figure 2.2 - Secured System with IDS</i> .....   | 22  |
| <i>Figure 2.3 - Common segments of the Intrusion Detection Framework.</i> .....   | 31  |
| <i>Figure 2.4 - Arrangement of an Intrusion Detection System inside an organisational system [56].</i> .....  | 33  |
| <i>Figure 3.1- A Typical Anomaly Detection System</i> .....   | 44  |
| <i>Figure 3.2 - A Conceptual Use of Neural Nets in Intrusion Detection [68]</i> .....   | 51  |
| <i>Figure 3.3 - A Trivial Bayesian Network Modelling Intrusive Activity.</i> .....  | 54  |
| <i>Figure 3.4 - A Generic Intrusion Detection Model.</i> .....  | 60  |
| <i>Figure 4.1- Machine Learning Algorithm, supported [84].</i> .....  | 71  |
| <i>Figure 4.2 - A linear two-class classification drawback [88].</i> .....  | 74  |
| <i>Figure 4.3 - A rectilinear regression drawback.</i> .....  | 75  |
| <i>Figure 4.4 - A group of labelled data points from the test data that are linearly separable. The data points have been separated by 4 hyper planes which will classify them correctly.</i> .....   | 89  |
| <i>Figure 4.5 - An optimum placement of the hyper plane that divides the 2 classes of the check knowledge whereas maximising the scale of the margin.</i> .....   | 90  |
| <i>Figure 4.6 - Left: A non-linearly severable knowledge set within the input area. Middle: an equivalent input in an exceedingly feature area wherever a linear classification is found. Right: The input within the feature area is then remodelled into the input space</i> .....    | 91  |
| <i>Figure 4.7 - Fascinating clusters might exist at many levels. Additionally to A, B and C, in addition the cluster D, that may be a combination of A and B, are fascinating of the subsequent steps.</i> .....  | 98  |
| <i>Figure 4.8 - Inter-cluster similarity outlined by single-link, complete-link, and average-link.</i> .....  | 101 |
| <i>Figure 4.9 - A example of Hidden Markov Model</i> .....  | 104 |
| <i>Figure 4.10 - Structure of SOM [8].</i> .....  | 105 |
| <i>Figure 4.11 - SOM Learning Example</i> .....   | 107 |
| <i>Figure 4.12 - Separate neighbourhoods (size 0,1 and 2) of the axis most unit: (a) hexangular lattice, (b) rectangular lattice. The intimate polygon figure corresponds to 0-neighbourhood, the instant to the 1-neighbourhood and also the largest to the 2- neighbourhood</i> ..... | 108 |
| <i>Figure 4.13 - Totally different map shapes. The default forms (a) and two shapes anywhere the map topology accommodates spherical data: cylinder (b) and toroid (c).</i> .....   | 109 |
| <i>Figure 4.14 - Change the least difficult matching unit (BMU) and its neighbours towards the data test with x. The strong and dabbed lines compare to situation before and when change, respectively.</i> .....   | 111 |
| <i>Figure 4.15 - Completely diverse neighbourhood capacities.</i> .....   | 112 |
| <i>Figure 4.16 - Dissimilar knowledge rate capacities.</i> .....  | 113 |
| <i>Figure 4.17 - Two feature effects created by the area work: (an) outskirts impact and (b) adding units.</i> .....  | 115 |
| <i>Figure 4.18 - SOM Classification Points in Thread Space</i> .....  | 118 |
| <i>Figure 4.19 - Component Planes for the Data Points in 3-D space.</i> .....   | 119 |
| <i>Figure 6.1 - System Architecture</i> .....   | 147 |
| <i>Figure 6.2 - Learning Phase of SOM</i> .....   | 155 |
| <i>Figure 6.3 - SOM Prepared with Conditional Probability Transformed Dataset</i> .....   | 172 |
| <i>Figure 6.4 - SOM Prepared with Indicator Variable Transformed Dataset</i> .....  | 173 |



## List of Tables

|   |     |
|---|-----|
| <i>Table 5.1</i> - KDD-CUP Centre of Attention .....  | 127 |
| <i>Table 5.2</i> - Features of KDD CUP99 .....  | 131 |
| <i>Table 5.3</i> - Basic features of individual TCP connections.....                        | 131 |
| <i>Table 5.4</i> - Content features within a connection suggested by domain knowledge ..... | 132 |
| <i>Table 5.5</i> - Traffic features computed using a two-second time window .....           | 132 |
| <i>Table 5.6</i> - Symbolic Features .....  | 133 |
| <i>Table 5.7</i> - Flag Features [1].....   | 135 |
| <i>Table 5.8</i> - Classification of Attacks.....   | 140 |
| <i>Table 5.9</i> - Effects of Different Kind of Attacks .....                               | 143 |
| <i>Table 6.1</i> - Sample Dataset for Indicator Variable Conversion .....                   | 151 |
| <i>Table 6.2</i> - Converted Dataset after Indicator Variable Conversion.....               | 151 |
| <i>Table 6.3</i> - Sample Dataset for Conditional Probability Conversion.....               | 153 |
| <i>Table 6.4</i> - Converted Dataset after Conditional Probability Conversion .....         | 153 |
| <i>Table 6.5</i> - Results for Detecting Attack Connections.....                            | 169 |
| <i>Table 6.6</i> - Results for Detecting Normal Connections .....                           | 170 |

## List of Abbreviations

|        |  |
|--------|--|
| IDS    | Intrusion Detection System                 |
| LSHSN  | Large-Scale High-Speed Networks            |
| KDD    | Knowledge Discovery in Data Competitions   |
| ITS    | Intrusion Tolerant System                  |
| NIDS   | Network Intrusion Detection System         |
| MADAM  | Mining Audit Data for Automated Model      |
| HIDS   | Host Based Intrusion Detection System      |
| IDES   | Intrusion Detection Expert System          |
| NIDES  | Network Intrusion Detection Expert System  |
| SVM    | Support Vector Machines                    |
| SOM    | Self Organizing Map                        |
| NB     | Naïve Bayes                                |
| ML     | Machine Learning                           |
| DARPA  | Defense Advanced Research Projects Agency  |
| HMM    | Hidden Markov Model                        |
| HTML   | Hyper Text Markup Language                 |
| DOS    | Denial-of- Service                         |
| U to R | User to Root Attacks                       |
| R to L | Remote to Local Attacks                    |
| DR     | Detection Rate                             |
| FPR    | False Positive Rate                        |
| WEKA   | Waikato Environment for Knowledge Analysis |

## Abstract

In this digital period, internet has turned into an indispensable wellspring of correspondence in just about every calling. With the expanded use of system engineering, its security has developed to be exceptionally discriminating issue as the workstations in distinctive association hold very private data and touchy information. The system used to screen the system security is known as Network detection. Intrusion detection is to get ambushes against a machine structure. It is a discriminating enhancement great to go part and additionally an element extent of examination. In Information Security, Intrusion recognizable proof is the showing of placing exercises that attempt to deal the protection, respectability or availability of a benefit. It accepts an astoundingly key part in waylay area, security check and framework inspect. One of the vital tests to Intrusion Detection is the issue of misjudgement, misdetection and unsuccessful deficiency of steady response to the strike. In the past years, as the second line of boundary after firewall, the Intrusion Detection strategy has got speedy progression.

This research work prepares two diverse Machine Learning techniques, both supervised and unsupervised, for Network Intrusion Detection. These techniques are Naïve Bayes (supervised learning) and Self Organizing Maps (unsupervised learning). The KDD Cup 99 dataset is utilized for Intrusion Detection Problem. As KDD Cup 99 dataset holds some symbolic attribute and also numeric attributes, two sorts of transformation technique have been utilized for these properties. These are conditional probabilities conversion technique and indicator variables transformation. The two machine learning procedures are prepared on both kind of transformed dataset and afterward their outcomes are looked at with respect to the correctness of intrusion detection.

**Keywords:** Network Intrusion Detection, supervised learning, unsupervised learning, Self-Organizing Map, Naïve Bayes, Conditional Probability Symbolic Conversion, Indicator Variable Symbolic Conversion, KDD Cup 1999.